Z Erziehungswiss (2017) 20:748–771
DOI 10.1007/s11618-017-0737-3

ZfE CrossMark

ALLGEMEINER TEIL

Die G8-Reform in Baden-Württemberg: Kompetenzen, Wohlbefinden und Freizeitverhalten vor und nach der Reform

Nicolas Hübner · Wolfgang Wagner · Jochen Kramer · Benjamin Nagengast · Ulrich Trautwein

Online publiziert: 15. März 2017

Zusammenfassung Die Konsequenzen der Einführung des achtjährigen Gymnasiums (G8) werden in Politik und Öffentlichkeit kontrovers diskutiert, u. a. weil es lange an belastbaren empirischen Daten mangelte. Der vorliegende Beitrag untersucht die Frage, ob sich Abiturientinnen und Abiturienten aus G8- und G9-Jahrgängen in Baden-Württemberg im Hinblick auf verschiedene Kompetenzbereiche sowie in ihren Selbstberichten zu ihrer schulischen Beanspruchung, ihren gesundheitlichen Beschwerden und in ihrem Freizeitverhalten unterschieden. Die Analysen beruhen auf Daten von vier Kohorten der Zusatzstudie Baden-Württemberg des Nationalen Bildungspanels: der letzte reine G9-Jahrgang (N = 1341), der G9-Doppeljahrgang (N = 1284), der G8-Doppeljahrgang (N = 1293) und der erste reine G8-Jahrgang (N = 1292). Im Hinblick auf die fachspezifischen Kompetenzen von Schülerinnen und Schülern zeigten sich zwischen G8- und G9-Jahrgängen in den Bereichen Mathematik und Physik keine Unterschiede, in Biologie geringfügige und in der Englisch-Lesekompetenz substanzielle Unterschiede zugunsten der Schülerinnen und Schüler

Die beiden erstgenannten Autoren haben in gleicher Weise bei der Entstehung des vorliegenden Beitrags mitgewirkt und werden in alphabetischer Reihenfolge aufgeführt. Anfragen zum Manuskript richten Sie bitte an Nicolas Hübner, Hector-Institut für Empirische Bildungsforschung, Europastraße 6, 72072 Tübingen. E-Mail: nicolas.huebner@uni-tuebingen.de.

N. Hübner (✉) · Dr. W. Wagner · Dr. J. Kramer · Prof. Dr. B. Nagengast · Prof. Dr. U. Trautwein
Hector-Institut für Empirische Bildungsforschung, Universität Tübingen,
Europastraße 6, 72072 Tübingen, Deutschland
E-Mail: nicolas.huebner@uni-tuebingen.de

Dr. W. Wagner
E-Mail: wolfgang.wagner@uni-tuebingen.de

Prof. Dr. B. Nagengast
E-Mail: benjamin.nagengast@uni-tuebingen.de

Prof. Dr. U. Trautwein
E-Mail: ulrich.trautwein@uni-tuebingen.de

aus G9-Jahrgängen. Bei der schulischen Beanspruchung und den gesundheitlichen Beschwerden fanden sich in G8-Jahrgängen substanziell höhere Werte. Im Hinblick auf das Freizeitverhalten fanden sich uneinheitliche Ergebnisse. Fragen nach Ursachen der Reformeffekte sowie Implikationen der Befunde für die Schulpolitik werden abschließend diskutiert.

Schlüsselwörter G8/G9 Reform · Kompetenzen · Schulisches Beanspruchungserleben · Gesundheitliche Beschwerden · Freizeitverhalten

The G8 reform in Baden-Württemberg: competencies, wellbeing and leisure time before and after the reform

Abstract The introduction of the eight-year high school stream was considered controversial by politicians and the public, partially because of the lack of empirical data supporting the decision. The present study compared students from G8 and G9 cohorts in Baden-Württemberg regarding cognitive variables such as competence in mathematics, English reading, physics and biology as well as non-cognitive outcomes such as school related stress, health problems and leisure time activities. Based on representative data from the National Educational Panel Study (NEPS; Add-on-Study Baden-Württemberg), students from four cohorts spanning 2011 to 2013 were compared. In regard to the subject-specific competences we found no differences between students from G8 and G9 cohorts in mathematics and physics, minor disadvantages for G8 students in biology and the largest disadvantage for G8 students in English reading achievement. Concerning stress and health problems we found disadvantages for G8 students, whereas effects for leisure time use remained inconsistent. Interpretations of the findings and possible implications are discussed.

Keywords Competencies · G8-reform · Health problems · Leisure activities · School related stress

1 Einleitung

Die Verkürzung der ursprünglich neunjährigen gymnasialen Schulzeit auf acht Jahre bei gleichzeitiger Beibehaltung des Gesamtvolumens von 265 Jahreswochenstunden wurde in zahlreichen westdeutschen Bundesländern in der ersten Dekade des neuen Millenniums umgesetzt (Trautwein und Neumann 2008; KMK 2014). Diese flächendeckende Einführung von G8 wurde und wird von Befürwortern und Gegnern kontrovers diskutiert (*Schul-Volksbegehren in Niedersachsen* 2011; Tulodetzki und Gohr 2012; Jacobsen und Buhse 2013; Vieth-Entus 2014). In Niedersachsen wurde mit Verweis auf die vermuteten negativen Effekte der G8-Reform inzwischen eine landesweite Rückkehr zu G9 zum Schuljahr 2015/2016 (KMK 2014; Kultusministerium Niedersachsen 2014) veranlasst, andere Bundesländer haben G9-Optionen eingeführt.

Die intensive öffentliche Diskussion um G8/G9 steht in auffälligem Kontrast zu einem „Schweigen“ der Erziehungswissenschaft, der nach Weiler (2003) sowohl bei

der Einführung von G8 als auch bei der jetzigen (partiellen) Rückkehr zu G9 keine bedeutsame Rolle zukam (für eine Ausnahme, vgl. Spiewak 2014). Tatsächlich lässt sich der derzeitige Forschungsstand zu den Reformeffekten der Schulzeitverkürzung als unbefriedigend bezeichnen (Kühn et al. 2013). Dies drückt sich ebenfalls im Fehlen eines konkreten theoretischen Rahmenmodells aus, welches die Reform z. B. in Bezug auf ihre Entstehung, ihre Ziele und potenziell wirksam werdenden Mechanismen oder Nebenwirkungen auf der Ebene des Unterrichts, der Schule oder unter Rückbezug auf weitere Akteure systematisch fundiert. In dem vorliegenden Beitrag werden beispielhaft für ein Bundesland Daten zu den Effekten von G8 zum Zeitpunkt des Abiturs vorgestellt und dazu genutzt, die Rolle von empirischen Befunden in der politischen Meinungsbildung zu diskutieren.

2 Diskussionen und Forschungsbefunde zu Schulzeitverkürzungen

Das Gymnasium und seine Weiterentwicklung haben schon immer in besonderer Weise die Aufmerksamkeit von Bildungspolitik und Öffentlichkeit gefunden (Fuchs 2004; Trautwein und Neumann 2008). Ein besonders umstrittenes Thema war und ist die Beschulungsdauer auf dem Gymnasium. Für die Einführung bzw. Beibehaltung von G8 (z. B. Herrmann 2002; Kühn et al. 2013) wurden u. a. ökonomische und demographische Argumente vorgetragen; darüber hinaus wurde auf Straffungsmöglichkeiten im Curriculum des G9 sowie eine wahrgenommene Entwicklungsakzeleration von Kindern und Jugendlichen verwiesen, weshalb G8 auch eine Stärkung der Eigenverantwortlichkeit junger Erwachsener ermögliche. Hingegen kritisieren Befürworter des G9 die Argumente für G8 als zu vereinfacht (vgl. Kühn et al. 2013; siehe auch Herrmann 2002). Besonders hervorgehoben wird dabei die Qualität gymnasialer Bildung, die durch G9 besser garantiert werden könne als durch G8, wobei neben Aspekten des Kompetenzerwerbs und des interessenorientierten Lernens auch mögliche positive Effekte auf die Persönlichkeitsentwicklung im weiteren Sinne genannt werden. Zusätzliche Argumente, die für G9 angeführt werden, betreffen negative Auswirkungen von G8 auf die Berufs- und Studienorientierung, Auslandsaufenthalte, extracurriculare Aktivitäten, Stresserleben und gesundheitliche Beschwerden. Zudem werden mögliche negative Effekte von G8 in leistungsheterogenen Klassen sowie in Hinblick auf die Durchlässigkeit des Schulsystems (im Sinne der Aufwärtsmobilität) thematisiert.

Insgesamt ist die empirische Datenlage im Vergleich zur Bedeutung der Thematik und zum Ausmaß der Umsetzung der flächendeckenden Reformmaßnahmen in fast allen Bundesländern eher dünn und fällt sehr viel weniger eindeutig aus als viele Befürworter von G8 oder G9 suggerieren. Man kann in dieser Debatte drei unterschiedliche Datenquellen unterscheiden (vgl. Kühn et al. 2013):

Erstens sind Befunde aus Studien mit begabten und hochbegabten Schülerinnen und Schülern zu nennen (z. B. Heller 2002). Die oftmals vorgetragenen positiven Befunde aus Studien zu verkürzten Schulzeiten für diese Schülerschaft („Hochbegabtenzüge“) eignen sich jedoch nicht für eine Generalisierung auf breitere Schülergruppen, von methodischen Problemen der entsprechenden Studien ganz abgesehen.

Zweitens werden teilweise internationale Befunde zum Zusammenhang von Beschulungsdauer und Schulleistungen in die Diskussion eingebracht. Inzwischen liegen eine Reihe von Reviews vor, die – bei relativ großer Streuung der Befunde – in der Mehrheit einen eher positiven Zusammenhang zwischen Beschulungsdauer und Schulleistung bzw. anderen kognitiven Kriteriumsmaßen nahelegen (vgl. Ceci 1991; Patall et al. 2010; Scheerens 2014). Allerdings unterscheiden sich die berichteten Studien im Hinblick auf Stichproben, Zeitmaße und Zielkriterien so stark, dass ihre Implikationen für die Situation in Deutschland nur sehr gering sind.

Die dritte Gruppe von Studien, Vergleiche von G8- und G9-Regelgymnasien, sind potenziell besonders aussagekräftig, allerdings ist die Datenlage in Hinblick auf relevante Kriteriumsmaße sehr begrenzt. Die vorliegenden Leistungsvergleiche zwischen Schülerinnen und Schülern aus G8- und G9-Systemen beziehen sich nahezu ausnahmslos auf Schulnoten. Hier zeigten sich überwiegend keine oder kleine Effekte teilweise gegensätzlicher Natur, die nur teilweise statistisch signifikant waren (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen 2013; Büttner und Thomsen 2013). Generell ist bei der Interpretation der Effekte der Reform auf Schulnoten kritisch anzumerken, dass sich Schulnoten nur begrenzt dafür eignen, Reformeffekte auf die Leistungsentwicklung adäquat abzubilden, da die Noten starken Referenzgruppeneffekten unterliegen können (vgl. Trautwein et al. 2006; Trautwein et al. 2008).

Auch für weitere Kriteriumsmaße wie Lernverhalten und Beanspruchungserleben ist die empirische Befundlage dünn und uneinheitlich. Böhm-Kasper und Weishaupt (2002) untersuchten in einer Studie verschiedene psychosoziale Merkmale wie beispielsweise den Leistungsdruck, das Schulklima, Beanspruchungsgefühle und die Konkurrenz zwischen Schülerinnen und Schülern in der Klassenstufe 8 und in der Kursstufe. Sie fanden uneinheitliche Effekte innerhalb und zwischen den untersuchten Bundesländern und deutliche Geschlechtereffekte. Unabhängig vom Bundesland fühlten sich Schülerinnen höher belastet als Schüler. Auch Milde-Busch et al. (2010) gingen der Frage nach Zusammenhängen einer verkürzten Gymnasialzeit mit dem gesundheitlichen Beschwerden bei Münchener Schülerinnen und Schülern der Klassestufen 10 (G8) und 11 (G9) nach und fanden lediglich im Hinblick auf den Anteil unverplanter Freizeit und in Bezug auf die Einschätzung der Erholung in dieser Zeit substanzielle Unterschiede zuungunsten der G8-Schülerinnen und Schüler. Quis (2015) untersuchte bereits Schülerinnen und Schüler des G8-G9-Doppeljahrgangs in Baden-Württemberg hinsichtlich möglicher Unterschiede im Wohlbefinden, ebenfalls auf Basis der Daten des Nationalen Bildungspanels, jedoch ohne den ersten reinen G8-Jahrgang. Es zeigte sich ein Unterschied von rund 30 % einer Standardabweichung beim Beanspruchungserleben und 10 % einer Standardabweichung bei den gesundheitlichen Beschwerden zuungunsten der G8-Kohorte. Trotz des Fehlens eines klaren konzeptuellen pädagogischen Rahmens für die Reform lassen sich natürlich mögliche Wirkfaktoren aus der wissenschaftlichen Literatur heranziehen. Im vorliegenden Falle liegen die Wirkfaktoren (Änderungen im Curriculum, Beibehaltung von Gesamtstundenzahl, Veränderung der Gesamt-Schulzeit, Alter beim Abitur, etc.) in einer komplexen Mischung vor, deren Gesamteffekt nur schwer zu antizipieren ist. Aus diesem Grund wollen wir unseren Artikel nicht im Sinne einer experimentellen Studie als Prüfung einer expliziten Theorie (z. B. in Bezug auf die

Lernzeit; Bloom 1968; Carroll 1963, 1989) bzw. eines bestimmten Wirkfaktors unter Kontrolle aller anderen verstehen, sondern wir untersuchen ein Maßnahmenpaket. Eine theoretische Einbettung der G8-Reform (auch über den Bezug zur Lernzeit hinaus) ist daher notwendig, kann von uns aber in diesem Artikel, der zunächst Ergebnisse der Reform darstellt, nicht vollumfänglich geleistet werden. Gleichwohl sei darauf verwiesen, dass aktuelle Studien mit Schülerinnen und Schüler aus G8- und G9-Jahrgängen auf Basis von PISA-Daten in Klassenstufe 9 kleine Vorsprünge der G8-Kohorte nahelegen, schwache Schüler nicht von der Reform zu profitieren scheinen und Leistungsunterschiede zwischen starken und schwachen Schülerinnen und Schüler sich verstärken (Huebener et al. 2016).

3 Umsetzung der G8-Reform in Baden-Württemberg

In der öffentlichen Wahrnehmung wird häufig nicht beachtet, dass die G8-Reformen in den einzelnen Bundesländern unterschiedlich implementiert wurden. Neben der „reinen“ Schulzeitverkürzung sollten deshalb immer auch weitere Faktoren identifiziert werden, die einen Effekt auf die Kriteriumsmaße haben können.

In Baden-Württemberg wurden im Zuge der Umsetzung der G8-Reform die durchschnittlichen Wochenstunden am allgemeinbildenden Gymnasium (Trautwein und Neumann 2008) erhöht, um die Vorgabe der Kultusministerkonferenz von 265 Jahreswochenstunden bis zum Abitur einzuhalten. Darüber sah der gemeinsam mit G8 eingeführte neue Bildungsplan für das Gymnasium die Einführung von Bildungsstandards mit Kerncurricula, die Verpflichtung zur Erstellung eines Schulcurriculums, das Erlernen einer zweiten Fremdsprache ab Klassenstufe 5 und die Einführung des Faches Naturwissenschaft und Technik (NwT) vor (Kultusministerium Baden-Württemberg 2004a, 2004b).

Ein Vergleich der Stundentafeln zeigt, dass sich als bedeutsamer Unterschied in Bezug auf die G8- und der G9-Systeme z. B. die Stundenreduktionen im Fach Mathematik in der Sekundarstufe I (G8: 24 h; G9: 28 h) nennen lässt. Im Fach Biologie erfolgte eine Reduktion um durchschnittlich 2 h in der Sekundarstufe I. Für die erste und zweite Fremdsprache kam es zu einer Stundenreduktion in der Sekundarstufe I, die durch die Einführung von acht Jahreswochenstunden Grundschulenglisch für alle Jahrgänge ab dem Einschulungsjahr 2004/2005 kompensiert wurde. Im Fach Physik blieb das Stundenvolumen gleich (Landesinstitut für Schulentwicklung 1999; Kultusministerium Baden-Württemberg 2004b). Die hier berücksichtigten G8-Jahrgänge hatten in der Grundschule also noch keinen Englischunterricht, wie von der Reform für aktuelle G8-Jahrgänge vorgesehen. Dies sollte bei einem Vergleich der Englischleistung von G8- und G9-Jahrgängen stets berücksichtigt werden.

4 Fragestellung

Welche Effekte die G8-Reformen in den einzelnen Bundesländern hatten, ist höchst umstritten und empirisch weitgehend ungeklärt. Für das Bundesland Baden-Württemberg werden in dieser Studie auf der Basis belastbarer Daten nun erstmals zen-

trale Kriteriumsmaße untersucht. Dabei ist zu beachten, dass – wie in den anderen Bundesländern auch – die G8-Reform in Baden-Württemberg von weiteren Maßnahmen begleitet wurde.

In der hier vorgestellten Studie wird die Veränderung in den Kompetenzen in vier Domänen (Mathematik, Englisch-Lesekompetenz, Physik und Biologie) untersucht. Hierbei stellt sich als zentrale Frage, ob und in welchem Maße sich die G8-Reform in geringeren Kompetenzen niederschlug. Im Hinblick auf das Freizeitverhalten wurde die Befürchtung geäußert, dass Abiturienten in G8 weniger Zeit für außerunterrichtliche Aktivitäten wie Sport und Musik haben könnten (z. B. Greiner und Himmelrath 2014; Laging et al. 2014). In der vorliegenden Studie konnten insgesamt elf Freizeitaktivitäten herangezogen werden, um etwaige Effekte zu prüfen. Schließlich wurde in Bezug auf das Beanspruchungserleben und die selbst eingeschätzten gesundheitlichen Beschwerden untersucht, ob sich diese zwischen den G8- und G9-Abiturienten unterscheiden.

5 Methode

5.1 Stichprobe

Es wurden Daten aus drei Erhebungswellen (Studiennummern: A72, A73 und A74) aus dem Scientific Use File Version 3.0.0.[1] der NEPS Zusatzstudie Baden-Württemberg (Blossfeld et al. 2011) herangezogen (vgl. Tab. 1). Konkret wurden der G9-Abschlussjahrgang 2011 (Welle I), der „Doppeljahrgang“ 2012 (Welle II) sowie der erste reine G8-Abschlussjahrgang 2013 (Welle III) erfasst. Es handelt sich also um ein Kohorten-Kontroll-Design (Shadish et al. 2002), welches hier die Grundlage für ein natürliches Experiment bildet (Murnane und Willett 2011).

Insgesamt nahmen 48 zufällig gezogene Schulen aus Baden-Württemberg (zwei dieser Schulen konnten aus organisatorischen Gründen in der ersten Welle nicht berücksichtigt werden) mit insgesamt rund 5000 Abiturienten (Welle 1: $N = 1341$; Welle 2: $N = 2577$; Welle 3: $N = 1292$) an der Untersuchung teil.[2]

[1] Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS), Zusatzstudie Baden-Württemberg, doi:10.5157/NEPS:BW:3.0.0. Die Daten des NEPS wurden von 2008 bis 2013 als Teil des Rahmenprogramms zur Förderung der empirischen Bildungsforschung erhoben, welches vom Bundesministerium für Bildung und Forschung (BMBF) finanziert wurde. Seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e. V. (LIfBi) an der Otto-Friedrich-Universität Bamberg in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt.

[2] In der ersten Welle liegen zusätzlich zu den Daten der G9-Schülerinnen und Schüler auch Daten von 52 Schülerinnen und Schülern aus dem G8-Schnellläufer Jahrgang vor. Diese wurden als Teil des ursprünglichen G9-Systems betrachtet (das sogenannte G8-Schnellläuferklassen umfasste) und bei den Analysen entsprechend als G9-Schüler kodiert. Der Ausschluss der G8-Schnellläufer hatte keinen Einfluss auf die Ergebnisse.

Tab. 1 Stichprobengrößen der drei Wellen differenziert nach G8- und G9-Anteilen

Kohorte	2011		2012		2013	Gesamt	
Jahrgang	G8*	G9	G8	G9	G8	G8	G9
Teilnahme Schüler	52	1289	1284	1293	1292	2628	2582
Teilnahme Schulen	46	46	48	48	48	48	48
Teilnahmequote	95,7		90,0		94,0	92,5	
Gesamt	1341		2577		1292	5210	

Die Erhebung der ersten Kohorte fand im Zeitraum vom 03. bis zum 18. Mai 2011 statt. Die zweite Erhebung erfolgte vom 23. April bis zum 22. Mai 2012. Die Erhebung der dritten Kohorte wurde schließlich zwischen dem 13. Mai und dem 12. Juni 2015 durchgeführt (IEA DPC 2013, 2014a, 2014b). G8*: G8-Schnellläufer Jahrgang

5.2 Instrumente

Mathematische Kompetenz. Aufgaben zur Messung der mathematischen Kompetenz basierten auf dem Konzept der *Mathematical Literacy*, das auch in PISA und den Nationalen Bildungsstandards verwendet wird (NEPS 2011). Hierbei werden vier Inhaltsbereiche unterschieden: *Quantität, Raum und Form, Veränderung und Beziehungen* sowie *Daten und Zufall*, die sich wiederum in sechs Komponenten mathematischer Denkprozesse unterscheiden lassen: technische Fertigkeiten einsetzen, modellieren, argumentieren, kommunizieren, repräsentieren und Probleme lösen. Im Mathematiktest wurden jeweils vier Items in den Bereichen *Quantität* und *Raum und Form* sowie jeweils sechs Items in den Bereichen *Veränderung und Beziehung* und *Daten und Zufall* administriert (Duchhardt 2015). Insgesamt wurden in der NEPS Zusatzstudie 21 Mathematikitems im Multiple Choice oder offenen Antwortformat administriert, für deren Bearbeitung 30 Min. Zeit zur Verfügung standen. Die Aufgaben orientieren sich in der Mehrzahl an den Inhalten der Mittelstufe.[3]

Englisch-Lesekompetenz. Zur Erfassung der Englisch-Lesekompetenz wurde auf am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) entwickelte Aufgaben zurückgegriffen (Rupp et al. 2008). Diese Aufgaben berücksichtigen einerseits die Bildungsstandards für das Fach Englisch, auf der anderen Seite orientieren sie sich am Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER; Europarat 2001). Im Englischtest wurden insgesamt fünf Items auf dem Niveau B1, vier Items auf dem Niveau B1/B2 und 16 Items auf dem Niveau B2 administriert. Darüber hinaus lagen acht Items auf dem C1 Niveau des GER vor. Insgesamt wurden 33 Aufgaben, die die Niveaustufen B1 bis C1 (selbständige bis kompetente Sprachverwendung) abdecken, administriert (21 Items pro Testheft). Die Bearbeitungszeit lag bei 30 Min. (Hübner et al. 2016b).

Biologische Kompetenz. Die Erfassung der – in der NEPS-Studie so bezeichneten – „biologischen Kompetenz“ erfolgte anhand eines im Rahmen der EVAMAR II-Studie (Eberle et al. 2008) entwickelten Instruments. Ähnlich wie bei der mathematischen Kompetenz wurde zunächst eine Unterteilung des Konstrukts in Inhaltsbe-

[3] Ein Item wurde im Zuge der Skalierung ausgeschlossen (vgl. Duchhardt 2015).

reiche und drei Klassen kognitiver Anforderungsbereiche vorgenommen. Im Biologietest wurden mit 27 Items die Bereiche *Cytologie, Anatomie und Soffwechsel*, mit 10 Items die Bereiche *Informationsverarbeitung, Verhalten und Immunbiologie* und mit 7 Items die Bereiche *Genetik und Entwicklungsbiologie* erfasst. Darüber hinaus wurden 11 Items zum Thema *Ökologie* sowie 5 Items im Bereich *Systematik und Evaluation* administriert.

Bei den kognitiven Anforderungsbereichen handelt es sich zunächst um die Stufe I, die sich mit dem Reproduzieren und Anwenden von Eingeübtem beschäftigt, und um Stufe II, die kognitive Operationen erfordert, die auf das Umstrukturieren und Übertragen von Inhalten abzielen. Die letzte Stufe III nimmt schließlich Operationen des Beurteilens und Problemlösens in den Fokus (vgl. NEPS 2011). In der NEPS-Zusatzstudie Baden-Württemberg wurden Biologische Kompetenzen mit insgesamt 60 Items gemessen. Jede Schülerin und jeder Schüler sollte im Rahmen des Booklet-Designs dabei ein Ausschnitt von 36 Items bearbeiten. Die vorgegebene Bearbeitungszeit betrug insgesamt 45 Min. Die Items wurden in Multiple Choice Formaten oder in offenen Antwortformaten präsentiert (NEPS 2011). Die Aufgaben orientieren sich primär an den Inhalten der Kursstufe (Hübner et al. 2016a).

Physikalische Kompetenz. Die physikalische Kompetenz wurde mit 41 Items erfasst, die zum Teil aus vorhandenen Instrumenten (z. B. TIMSS; Baumert et al. 1999) übernommen und zum Teil speziell für die beiden NEPS-Zusatzstudien (Thüringen, Baden-Württemberg) entwickelt wurden (NEPS 2011)[4]. Hierbei sollte jede Schülerin und jeder Schüler einen Ausschnitt aller Items (19 bis 21 Items pro Testheft) bearbeiten. Im Physiktest wurden drei Items aus dem Bereich Elektrische Felder und Wechselwirkung, sechs Items aus dem Bereich Magnetische Felder und Elektromagnetische Induktion und zwei Items aus dem Bereich Spezielle Relativitätstheorie administriert. Darüber hinaus beinhaltete der Test jeweils vier Items aus den Bereichen Wellen, Quantenphysik: Quanten und Materie, Dynamik: Schwingungen und Dynamik: Mechanik des starren Körpers. Zuletzt wurden für die Bereiche Optik und Thermodynamik jeweils sieben Items administriert. Die Bearbeitungszeit für den Test lag ebenfalls bei insgesamt 45 min. Die Items waren im Multiple Choice, Forced Choice sowie im offenen Antwortformat formuliert. Die Konstruktion dieser Items orientiert sich an den Einheitlichen Prüfungsanforderungen für die Abiturprüfung (EPA) in Physik. Die Aufgaben orientieren sich primär an den Inhalten der Kursstufe (Hübner et al. 2016c).

Die Analyse aller Kompetenzen erfolgte simultan unter Verwendung eines vierdimensionalen Mehrgruppen-1PL-IRT-Modells. Für alle Tests zeigten sich substantielle Zusammenhänge zwischen der jeweiligen Note im Fach am Ende der Sekundarstufe II und der latenten Variable der Testleistung, die für Mathematik bei r = 0,59 lag, für Englisch bei r = 0,57, für Biologie bei r = 0,49 und für Physik bei r = 0,51. Die Kodierung der Items aller Kompetenztests in „korrekt“ und „falsch“ liegt im aktuellen Scientific Use File 3.0.0 (Blossfeld et al. 2011) bereits vor, so-

[4] Wir bedanken uns im Namen der Etappe 5 (Gymnasiale Oberstufe und Übergänge in (Fach-)Hochschule, Ausbildung oder Arbeitsmarkt) des NEPS für die Unterstützung bei der Erstellung des Physiktests bei Knut Neumann, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN).

dass dies nicht im Rahmen der Analysen der vorliegenden Studie erfolgte. Offene Antworten wurden entweder als „falsch“ (0) oder „korrekt“ (1) kodiert (in wenigen Fällen wurde auch ein dreistufiges Format mit der Kodierung „teilrichtig“ verwendet). Bei Items, bei denen sowohl ein numerischer Wert als auch eine Maßeinheit angegeben werden musste, wurden Antworten nur als korrekt gewertet, wenn beide Angaben richtig waren. Fehlende Angaben wurden gemäß der NEPS Standards (Pohl und Carstensen 2012) mit dem speziellen Missing Code „nicht bearbeitet“ gekennzeichnet, unleserlichen Angaben wurden als „nicht valide“ kodiert. Die Kodierung dieser offenen Items erfolgte computerbasiert per Syntax nach den Vorgaben des Auswertungsmanuals.

Schulbezogenes Beanspruchungserleben. Das schulbezogene Beanspruchungserleben wurde mit einer im Rahmen der NEPS-Zusatzstudie konzipierten Skala mit 15 Items erfasst. Dabei sollten die Abiturientinnen und Abiturienten die Zustimmung zu verschiedenen schulbezogenen Aussagen von 1 (stimme gar nicht zu) bis 4 (stimme völlig zu) beurteilen (Itembeispiele: „Wenn ich von der Schule nach Hause komme, bin ich angespannt“ und „Manchmal kann ich schwer einschlafen, weil mir Probleme aus der Schule durch den Kopf gehen“). Die interne Konsistenz der Skala (Cronbachs α) lag bei 0,91.

Gesundheitliche Beschwerden. Selbstberichte über gesundheitliche Beschwerden wurden mit insgesamt 26 Items erfasst. Hierbei handelt es sich um eine Skala, die bereits im Rahmen von PISA 2003 eingesetzt wurde (Bergmüller 2003). Schülerinnen und Schüler sollten jeweils die Häufigkeit des Auftretens verschiedener physischer und psychischer Symptome in den letzten sechs Wochen auf einer Skala von 1 (nie) bis 4 (öfter als sechsmal) angeben. Gefragt wurde hierbei beispielsweise nach „starkem Herzklopfen“, „Angst, dass alles zu viel wird“ oder „Erbrechen“. Die Skala wies eine interne Konsistenz (Cronbachs α) von 0,93 auf. Die Auswertungen der gesundheitlichen Beschwerden erfolgte simultan mit dem Beanspruchungserleben unter Anwendung von Graded Response Modellen (Samejima 1997; Embretson und Reise 2000).

Freizeitverhalten. Das Freizeitverhalten wurde in Stunden pro Woche für insgesamt elf Bereiche erfasst (Trautwein et al. 2010). Diese sind „Freizeitangebote der Schule (z. B. Sport-, Hobby-, Arbeitsgruppen)“, „Computer spielen, chatten etc.“, „Freundinnen und Freunde treffen“, „Fernsehen“, „Lesen“, „etwas mit der Familie unternehmen“, „Sport treiben (alleine, mit Freundinnen oder Freunden, im Verein)“, „zum Orchester, Kirchengruppen oder anderen Gruppen (außer Sport) gehen“, „Zeit mit anderen Hobbys verbringen (z. B. Instrumente, Basteln)“, „Nebenjob“ und „Sonstiges“. Diese Items sollten in Bezug auf die wöchentliche Beschäftigung und in Stunden beantwortet werden. Da es bei dieser Skala keine Antwortmöglichkeit für „keine Betätigung“ im jeweiligen Freizeitbereich gab, gilt zu beachten, dass fehlende Werte und „keine Betätigung“ bei diesen Skalen nicht eindeutig unterscheidbar sind und lediglich Aussagen über die relative Betätigungszeit von Schülerinnen und Schülern möglich ist, die eine konkrete Betätigungszeit angaben.

Sozialer und kultureller Hintergrund. Die Erfassung des sozialen Status der Familie der Schülerinnen und Schüler erfolgte auf Basis des International Socio-Economic Index of Occupational Status 2008 (ISEI-08; Ganzeboom und Treiman 2003). Aus dem ISEI-08 wurde in den vorliegenden Analysen der höchste ISEI (HISEI) aus dem jeweils höchsten ISEI der beiden Eltern gebildet. Der häusliche Buchbestand diente als Indikator des kulturellen Kapitals. Der familiäre Migrationsstatus wurde auf Basis des Geburtslands der Eltern bestimmt. Dabei wurde als Kriterium die Geburt mindestens eines Elternteils im Ausland festgelegt.

Kognitive Grundfähigkeiten. In der Zusatzstudie Baden-Württemberg wurden als nonverbale kognitive Grundfähigkeiten einerseits die Wahrnehmungsgeschwindigkeit und andererseits das schlussfolgernde Denken der Schülerinnen und Schüler erfasst (NEPS 2011). Konkret wurde die Wahrnehmungsgeschwindigkeit über den Bilder-Zeichen-Test (NEPS-BZT) erfasst, einem Test mit insgesamt 93 Items, für die in jeweils drei Itemblöcken mit 31 Items eine Bearbeitungszeit von jeweils 30 s pro Block vorgesehen war. Das schlussfolgernde Denken wurde mit Hilfe eines Matrizentests erfasst (NEPS-MAT), bei dem insgesamt 12 Items verwendet wurden. Der Test misst figurale kognitive Fähigkeiten (Brunner et al. 2014). Die Bearbeitung dieser Items erfolgte in drei Blöcken mit jeweils vier Items; hierfür standen jeweils drei Minuten Zeit pro Block zur Verfügung.

Kursbelegung. Im Rahmen des Schülerfragebogens wurde erfasst, ob die Schülerinnen und Schüler die Fächer Englisch, Biologie und Physik in der Oberstufe abgewählt bzw. als Kernfach gewählt hatten.

5.3 Statistische Analyse

Zunächst wurden Unterschiede in den Kursbelegungsquoten der verschiedenen Gruppen für die Bereiche Englisch-Lesekompetenz, Biologie und Physik mittels multinomialer logistischer Regressionen mit der Gruppenzugehörigkeit (G9W1, G9W2, G8W2, G8W3) als abhängiger Variable und dummy-kodierten Prädiktoren für Kurswahl (Kernfach bzw. Abwahl) anhand von Wald-Tests geprüft (Annahme: sämtliche Regressionskoeffizienten sind gleich Null). Zur adäquaten Untersuchung möglicher Unterschiede von Schülerinnen und Schülern aus G8- und G9-Jahrgängen wurde ein mehrstufiges Vorgehen gewählt. Die einzelnen Kompetenzmaße wurden zunächst mit einem eindimensionalen Rasch-Modell, bzw. Partial-Credit-Modell skaliert, um die psychometrische Qualität des Tests und der einzelnen Items zu überprüfen. Es wurden *Differential Item Functioning* (DIF)-Analysen für den HISEI, den Migrationshintergrund, das Geschlecht, das Kursniveau (bei Biologie- bzw. Physiktest) und die Erhebungswelle durchgeführt. Hierbei zeigte sich insgesamt nur auf wenigen Items starker DIF ≥ 0.60 Logits. Der Ausschluss dieser Items aus den Analysen führte zu keiner substantiellen Veränderung der Ergebnisse. DIF bedeutet nicht zwangsläufig, dass Items „unfair“ sind (item bias), sondern können auch ein Hinweis auf valide Unterschiede zwischen Subgruppen darstellen (Zumbo 1999). In verschiedenen Studien konnte darüber hinaus gezeigt werden, dass IRT-basierte Modellinferenzen bei moderaten Verletzungen der

Messinvarianzannahme relativ robust sind (Rupp und Zumbo 2006). Anschließend wurden die Kompetenzdaten, ebenso wie das schulische Beanspruchungserleben und die gesundheitlichen Beschwerden unter Verwendung von mehrdimensionalen Mehrgruppen-IRT-Modellen ausgewertet. Hierbei wurden zunächst vier latente Variablen (eine latente Dimension pro Kompetenzbereich) spezifiziert und deren Kovarianz frei geschätzt. Die Schätzung der latenten Mittelwerte der jeweiligen Kompetenzdimension erfolgte unter Verwendung des MLR-Schätzers. Indikatoren der latenten Dimensionen wurden als kategorial definiert. Der Vorteil einer mehrdimensionalen Skalierung gegenüber einer eindimensionalen Skalierung liegt in der theoretisch plausiblen und komplexen Abhängigkeit der Kompetenzen untereinander, die in einem mehrdimensionalen Modell explizit berücksichtigt werden kann (vgl. Reckase 2009) und mit einer höheren Teststärke einhergehen sollte. Die Analysen zum Freizeitverhalten erfolgten schließlich unter Verwendung von Mehrgruppen-Analysen für metrische Daten. Sämtliche Analysen wurden in Mplus 7.4 durchgeführt (Muthén und Muthén 1998–2012). Zu berücksichtigen ist, dass die von uns spezifizierten latenten Variablenmodelle Parameterschätzungen bezogen auf „messfehlerfrei“ erfasste Konstrukte ermöglichen (sofern die Modelle angemessen spezifiziert wurden), wobei geringere Reliabilität der Instrumente sich lediglich in (etwas) größeren Standardfehlern der Schätzungen niederschlägt.

Aus den jeweiligen Analysen resultierten Parameterschätzungen getrennt für vier Kohorten: Dem letzten reinen G9-Jahrgang (G9W1), dem G9-Doppeljahrgang (G9W2), dem G8-Doppeljahrgang (G8W2) und dem ersten reinen G8-Jahrgang (G8W3). Die Analyse der Reformeffekte erfolgte auf Basis verschiedener möglicher Kohortenvergleiche. Hierbei erfolgte sowohl ein Vergleich des Doppeljahrgangs, der beiden reinen Jahrgänge als auch ein Vergleich der gesamten G8- versus G9-Schülerinnen und Schüler. Diese Vergleiche bieten sich an, da sich Schülerinnen und Schüler im Doppeljahrgang zumindest theoretisch von Schülerinnen und Schülern in den beiden reinen Jahrgängen unterscheiden können. Hierbei gilt zu berücksichtigen, dass es einerseits keine Unterschiede zwischen Schülerinnen und Schülerinnen im Doppeljahrgang in Bezug auf den Lernstoff in der Oberstufe gab, während sich andererseits Schülerinnen und Schüler der reinen Jahrgänge diesbezüglich theoretisch unterscheiden können. Gleichzeitig führt die gemeinsame Unterrichtung von G8- und G9-Schülerinnen im Doppeljahrgang möglicherweise auch zu Referenzgruppeneffekten, die bei einem Vergleich der beiden reinen Jahrgänge praktisch auszuschließen sind. Aufgrund des neuen Curriculums in der Oberstufe, das beim ersten reinen G8-Jahrgang erstmals implementiert war, sollten sich die diesbezüglichen Befunde auch eher auf aktuelle G8-Jahrgänge generalisieren lassen.

Eine zentrale Herausforderung in quasi-experimentellen Designs besteht in der Trennung von Selektions- und Behandlungseffekten (Morgan und Winship 2007; Murnane und Willett 2011). Dazu wurden, unter Berücksichtigung zusätzlicher Daten des statistischen Landesamts, mögliche Selektionsunterschiede (z. B. in Übergangsquoten und Nichtversetztenquoten) geprüft. Anschließend wurde die Vergleichbarkeit der Schülerinnen und Schüler aus den G8- bzw. G9-Kohorten im Abschlussjahr bezüglich relevanter Hintergrundmerkmale untersucht. In einem letzten Schritt erfolgte schließlich die Untersuchung von Unterschieden der G8-

und G9-Schülerinnen und Schüler auf den Kriteriumsmaßen unter Kontrolle von Hintergrundmerkmalen.

Alle Analysen erfolgten zunächst im Rahmen eines unadjustierten Modells (ohne Kovariaten) und anschließend mit Adjustierung. Bei den berücksichtigten Kovariaten handelte es sich um das Geschlecht, den Migrationshintergrund, den häuslichen Buchbestand, kognitive Grundfähigkeiten[5] und Informationen zu Klassenwiederholungen in der gesamten Sekundarstufe. Die im Rahmen der unadjustierten Modellschätzungen aufgeführten Werte spiegeln die Mittelwerte der Variablen für die jeweiligen Kohorten ohne Adjustierung wider. Bei den Analysen der Kompetenzen wurde in einem zusätzlichen Modell das Kursniveau statistisch kontrolliert. Die adjustierten Modelle bieten zusätzlich zu den Modellen ohne Kovariaten die Möglichkeit einer Betrachtung von Unterschieden zwischen den jeweiligen Kohorten unter statistischer Kontrolle möglicher bestehender Gruppenunterschiede.

In den adjustierten Modellen wurden die Kovariaten vor der Analyse am Gesamtmittelwert über die drei Erhebungswellen zentriert. Unterscheiden sich die gruppenspezifischen Mittelwerte für die Kovariaten bei Regressionsgewichten ungleich Null, repräsentieren die Intercepts aus diesen Modellen adjustierte Gruppenmittelwerte für die „typische“ Schülerkomposition (als durchschnittliche Zusammensetzung in allen drei Erhebungswellen). Um Unterschiede zwischen Schülerinnen und Schülern aus den verschiedenen Kohorten auf den abhängigen Variablen zu untersuchen, wurden entsprechend Mittelwert- (unadjustierte Modelle) bzw. Intercept-Differenzwerte (adjustierte Modelle) inklusive Standardfehler geschätzt. Zur besseren Interpretierbarkeit möglicher Unterschiede wurden die aus den Analysen resultierenden Parameter linear transformiert. Für die Kompetenzen erfolgte eine Transformation auf eine Metrik mit $M = 500$ und $SD = 100$. Die Ergebnisse zum Wohlbefinden wurden in die T-Metrik überführt ($M = 50$ und $SD = 10$).

Die mit der Adjustierung verbundenen Annahmen sind zwar plausibel, müssen aber keinesfalls zwingend zu korrekten (oder wenigstens korrekteren) Schätzungen führen. So könnten etwa Kurswahlunterschiede auch als Reformeffekte interpretiert werden, sodass eine Adjustierung für das Kursniveau – je nach Blickwinkel – auch als ungerechtfertigt betrachtet werden kann. Auch eine Adjustierung auf Basis von im Abschlussjahrgang erhobener Maße der allgemeinen kognitiven Fähigkeiten kann prinzipiell zu Verzerrungen führen, da diese möglicherweise durch die Reform beeinflusst wurden. Aufgrund dieser Einschränkungen lässt sich kein klar zu favorisierendes Analysemodell formulieren, wenngleich der Einbezug von Hintergrundmerkmalen für die Schätzung unverzerrter Reformeffekte sinnvoll erscheint. Unterscheiden sich die Schätzungen aus verschiedenen Modellen (unterschiedliche Adjustierungen oder ohne Adjustierung) nur geringfügig, so kann dies im Sinne einer hohen „Robustheit“ der Befunde interpretiert werden.

Die Besonderheiten des Sampling Designs (Ziehung von Schulen, Surveygewichte; Schönberger und Aßmann 2014) wurden anhand entsprechender Mplus-Optionen berücksichtigt (Type = Complex; Weight-Option). Fehlende Werte wurden in den

[5] Die kognitiven Grundfähigkeiten hatten in allen Modellen einen nicht signifikanten oder einen geringen Einfluss auf die Ergebnisse. Die adjustierten Modelle mit und ohne Kontrolle der kognitiven Fähigkeiten unterschieden sich hinsichtlich ihrer Ergebnisse nicht bedeutsam voneinander.

Tab. 2 Deskriptive Statistik

	G9W1	G9W2	G8W2	G8W3
Durchschnittsnote Abitur	2,33 (0,62)	2,35 (0,61)	2,38 (0,61)	2,36 (0,64)
Alter	19,02 (0,60)	19,05 (0,51)	17,97 (0,39)	18,07 (0,62)
Weiblich	705 (55,1 %)	645 (54,7 %)	673 (55,6 %)	670 (55,4 %)
Migrationshintergrund	287 (22,5 %)	274 (23,4 %)	248 (20,1 %)	277 (22,9 %)
Sozioökonomischer Status	61,94 (19,31)	61,10 (19,20)	61,45 (19,55)	63,09 (18,17)
Schlussfolgerndes Denken	10,80 (1,23)	10,81 (1,28)	10,72 (1,26)	10,70 (1,30)
Wahrnehmungsgeschwindigkeit	65,85 (11,52)	64,87 (11,21)	64,55 (12,06)	65,37 (11,94)
Klassenwiederholung	131 (10,2 %)	112 (9,5 %)	19 (1,6 %)	130 (10,7 %)

Aufgeführt sind die Mittelwerte und Standardabweichungen bzw. für „weiblich“, „Migrationshintergrund“ und „Klassenwiederholung“ die absoluten und relativen Häufigkeiten; der sozioökonomische Status wurde mit dem höchsten ISEI gemessen

vorliegenden Analysen mithilfe der Full Information Maximum Likelihood-Methode (FIML) berücksichtigt.

6 Ergebnisse

6.1 Selektivitätsanalysen

Um zu prüfen, ob die Schülerschaft der vier berücksichtigten Kohorten vergleichbar war oder sich von vornherein (z. B. durch Klassenwiederholungen oder Schulwechsel) unterschied, wurden zunächst Hinweise auf unterschiedliche Selektionsprozesse näher untersucht. Auf Basis von Daten des Statistischen Landesamts Baden-Württemberg (2014b) wurden zunächst gymnasiale Übergangsquoten untersucht. Hierbei zeigte sich für die Jahre 2003, 2004 und 2005 ein geringfügiger Anstieg (2003: 35,3 %; 2004: 36,1 %; 2005: 37,8 %), der vor dem Hintergrund eines allgemein zunehmenden gymnasialen Übergangsverhaltens interpretiert werden kann.

Neben dem Übergangsverhalten sind auch die Anteile der Nichtversetzten und der Klassenwiederholer zentral für die Vergleichbarkeit von Schülerinnen und Schülern unterschiedlicher Kohorten. Die Nichtversetztenquote variierte zwischen Klassenstufe 5 und 11 nur geringfügig zwischen G8- und G9-Jahrgängen (G9: 0,4–3,1 %; G8: 0,4–3,7 %; (Schwarz-Jung 2008; Statistisches Landesamt Baden-Württemberg 2014a). Die Gruppe der G8-Schülerinnen und Schüler aus dem Doppeljahrgang wies allerdings einen besonders geringen Anteil an Klassenwiederholern auf (Statistische Ämter des Bundes und der Länder 2015). Da sich die Nichtversetztenquote aus Klassenwiederholern und Abgängern zusammensetzt, lässt sich daraus schließen, dass ein größerer Anteil der nichtversetzten Schülerinnen und Schüler aus dem letzten G9-Jahrgang eher auf eine andere Schulform wechselte, anstatt eine Klasse zu wiederholen. In der zweiten G8-Kohorte zeigte sich dann wieder eine Wiederholerquote vergleichbar mit der vor der Reform. Zur Vergleichbarkeit der Kohorten wurden Klassenwiederholungen daher in den adjustierten Analysen statistisch kontrolliert. Bei der Überprüfung möglicher Unterschiede in den Belegungsquoten zeigten sich für die Bereiche Physik ($\chi^2(6) = 5{,}68$, $p = 0{,}46$) und Biologie ($\chi^2(6) = 9{,}62$ $p = 0{,}14$)

keine Unterschiede. Für das Fach Englisch fand sich ein statistisch bedeutsamer Unterschied ($\chi^2(6) = 27{,}57$, $p < 0{,}001$). So wählten G9W1-Schülerinnen und Schüler Englisch weniger häufig als Kernfach (91 %; in den nachfolgenden Kohorten jeweils mindestens 94 %) und häufiger als Grundkurs (4 %; in den nachfolgenden Kohorten jeweils weniger als 1 %). Die Abwählerquote lag in sämtlichen Kohorten relativ konstant im Bereich von 5 bis 6 %.

Bei der deskriptiven Statistik (vgl. Tab. 2) zeigten sich in Bezug auf die meisten Variablen lediglich geringfügige Unterschiede zwischen den untersuchten Kohorten. Schülerinnen und Schüler aus G8-Kohorten waren im Mittel erwartungsgemäß ein Jahr jünger als Schülerinnen und Schüler aus G9-Kohorten.

6.2 Kompetenzen der Abiturientinnen und Abiturienten vor und nach der Oberstufenreform

Für die Mathematik ergaben sich in den adjustierten Modellen (ohne bzw. mit Kontrolle des Kursniveaus) keine statistisch signifikanten Unterschiede zwischen beiden G9- und G8-Kohorten (adjustiert ohne Kursniveau: M_{G9ges}-M_{G8ges}: –3, $p = 0{,}54$; adjustiert mit Kursniveau: M_{G9ges}-M_{G8ges}: –4, $p = 0{,}25$, siehe Tab. 3). Auch die übrigen Gruppenvergleiche in den Modellen mit Adjustierung waren nicht statistisch signifikant. Das Ergebnismuster des adjustierten Modells zeigte sich auch in den Modellen ohne Berücksichtigung weiterer Kovariaten.

Bei der Englisch-Lesekompetenz fanden sich statistisch signifikante Unterschiede zwischen beiden G9- und G8-Kohorten sowohl im adjustierten Modell ohne und unter Kontrolle des Kursniveaus. Im Mittel schnitten hier Schülerinnen und Schüler aus G9-Jahrgängen rund 18 bzw. 20 Punkte besser ab als Schülerinnen und Schüler aus G8-Jahrgängen. Gleiches gilt für die Unterschiede zwischen den Kohorten des Doppeljahrgangs und den beiden reinen G8- bzw. G9-Jahrgängen, bei denen ebenfalls jeweils die G9-Jahrgänge höhere Werte aufwiesen (vgl. Tab. 3).

Für die Biologische Kompetenz ergab sich ein Unterschied zwischen Schülerinnen und Schülern aus G9- und G8-Jahrgängen, der jedoch nur im adjustierten Modell mit Kursniveau statistisch signifikant war. Darüber hinaus unterschieden sich Schülerinnen und Schüler aus dem Doppeljahrgang in ihrer Biologiekompetenz nicht voneinander. Der Vergleich der reinen G9- bzw. G8-Jahrgänge ergab einen statistisch signifikanten Unterschied zugunsten der Schülerinnen und Schüler im letzten reinen G9-Jahrgang. Für die Physikkompetenz fanden sich ähnlich wie bei der Mathematikkompetenz keine Unterschiede zwischen Schülerinnen und Schülern aus G9- und G8-Jahrgängen (vgl. Tab. 3).[6]

6.3 Schulisches Beanspruchungserleben und gesundheitliche Beschwerden

Beim schulischen Beanspruchungserleben zeigte sich zunächst ein signifikanter Effekt für den Unterschied zwischen Schülerinnen und Schülern aus G9- und G8-Jahrgängen (M_{G9ges}-M_{G8ges}: –4,0, $p < 0{,}01$), bei dem G8-Schülerinenn und Schüler

[6] Die adjustierten Modelle mit und ohne Kontrolle der kognitiven Fähigkeiten unterschieden sich hinsichtlich ihrer Ergebnisse nicht bedeutsam voneinander.

Tab. 3 Adjustierte und unadjustierte Mittelwerte der fachspezifischen Kompetenzen Mathematik, Englisch-Lesekompetenz, Biologie und Physik für die jeweiligen Kohorten

Unadjustierte Ergebnisse						
	$G9W1_a$	$G9W2_b$	$G8W2_c$	$G8W3_d$	$G9_{ges}$–$G8_{ges}$	*p*
Mathematik	502	495 (5,50)	501 (5,08)	502 (6,02)	–2–2	0,54
Englisch	511_d	$509\ (5,18)_c$	$488\ (6,59)_b$	$493\ (5,27)_a$	20	<0,01
Biologie	507_d	500 (4,71)	499 (5,01)	$494\ (6,07)_a$	7	0,09
Physik	501	495 (5,79)	501 (6,05)	503 (5,26)	–4	0,25
Adjustierte Ergebnisse ohne Kursniveau						
Mathematik	502	496 (4,77)	498 (4,77)	504 (5,19)	–2	0,54
Englisch	511_d	$509\ (4,83)_c$	$486\ (6,23)_b$	$493\ (5,18)_a$	21	<0,01
Biologie	507_d	500 (4,25)	498 (5,01)	$495\ (5,47)_a$	7	0,07
Physik	501	496 (5,13)	497 (5,00)	506 (4,73)	–3	0,31
Adjustiert Ergebnisse mit Kursniveau						
Mathematik	500	496 (4,05)	498 (5,08)	506 (5,08)	–4	0,25
Englisch	510_d	$509\ (4,83)_c$	$489\ (5,97)_d$	$493\ (5,01)_a$	18	<0,01
Biologie	508_d	500 (4,40)	494 (5,31)	$498\ (4,86)_a$	8	0,04
Physik	500	499 (5,13)	498 (5,13)	503 (4,47)	–1	0,84

G9W1: Schülerinnen und Schüler aus G9-Jahrgängen der ersten Erhebungswelle; G9W2: Schülerinnen und Schüler aus G9-Jahrgängen der zweiten Erhebungswelle; G8W2: Schülerinnen und Schüler aus G8-Jahrgängen der zweiten Erhebungswelle; G8W3: Schülerinnen und Schüler aus G8-Jahrgängen der dritten Erhebungswelle; G9ges: Schülerinnen und Schüler aller G9-Kohorten; G8ges: Schülerinnen und Schüler aller G8-Kohorten; $G9_{ges}$–$G8_{ges}$: Mittelwertdifferenz aller G9- und G8-Jahrgänge. Die Metrik der latenten Variablen wurde transformiert auf $M = 500$ und $SD = 100$ auf Basis der gepoolten Mittelwerte bzw. Standardabweichungen. Die berichteten *p*-Werte beziehen sich auf zweiseitige Tests. Subskripte indizieren Unterschiede zwischen den jeweiligen Gruppen. Für Englisch waren alle Unterschiede signifikant mit $p < 0{,}001$, für Biologie mit $p < 0{,}05$. Es werden nur Unterschiede zwischen den Doppeljahrgängen, den reinen Jahrgängen und Unterschiede innerhalb der G8- und G9-Jahrgänge dargestellt. Die finale Analysestichprobe basiert auf Daten von $N = 4893$ Schülerinnen und Schülern, die zumindest am Schülerfragebogen oder an einem Leistungstest teilgenommen haben.

angaben, sich im Mittel höher beansprucht zu fühlen. Darüber hinaus fanden sich Unterschiede zwischen dem G8-G9-Doppeljahrgang (M_{G9W2}-M_{G8W2}: –3,1, $p < 0{,}01$) und bei einem Vergleich des letzten reinen G9-Jahrgangs mit dem ersten reinen G8-Jahrgang (M_{G9W1}-M_{G8W3}: –4,9, $p < 0{,}01$). Diese Ergebnisse waren äquivalent zu den Ergebnissen im unadjustierten Modell (vgl. Tab. 4).

In Bezug auf die gesundheitlichen Beschwerden zeigten sich im Mittel ebenfalls höhere Werte bei Schülerinnen und Schülern aus G8-Jahrgängen. Der Unterschied zwischen den Kohorten innerhalb des G8-G9-Doppeljahrgangs wurde nicht signifikant, wohingegen der Unterschied zwischen den beiden reinen Jahrgängen statistisch signifikant war. Es fanden sich ebenfalls keine Unterschiede zwischen diesen Ergebnissen und den Ergebnissen im unadjustierten Modell (vgl. Tab. 4).

6.4 Freizeitverhalten

Bei der Analyse der Angaben zu Zeitinvestitionen für Freizeitbereiche zeigten sich in vier der elf untersuchten Bereiche signifikante Unterschiede im adjustierten und im unadjustierten Modell zwischen G9- und G8-Jahrgängen (vgl. Tab. 5). Zu be-

Tab. 4 Adjustierte und unadjustierte Mittelwerte des schulischen Beanspruchungserleben und der wahrgenommenen gesundheitlichen Beschwerden nach Kohorte

Unadjustierte Ergebnisse						
	$G9W1_a$	$G9W2_b$	$G8W2_c$	$G8W3_d$	$G9_{ges}$–$G8_{ges}$	p
Beanspruchungserleben	$47{,}2_{bd}$	$49{,}0$ $(0{,}48)_{ac}$	$51{,}7$ $(0{,}55)_{bd}$	$52{,}1$ $(0{,}50)_{ac}$	–3,9	<0,01
Gesundheitliche Beschwerden	$48{,}2_{bd}$	$49{,}7$ $(0{,}43)_{a}$	$50{,}5$ $(0{,}42)_{d}$	$51{,}6$ $(0{,}44)_{ac}$	–2,1	<0,01
Adjustierte Ergebnisse ohne Kursniveau						
Beanspruchungserleben	$47{,}2_{bd}$	$48{,}9$ $(0{,}52)_{ac}$	$51{,}9$ $(0{,}45)_{bd}$	$52{,}0$ $(0{,}47)_{ac}$	–4,0	<0,01
Gesundheitliche Beschwerden	$48{,}3_{bd}$	$49{,}6$ $(0{,}57)_{a}$	$50{,}6$ $(0{,}64)_{d}$	$51{,}5$ $(0{,}67)_{ac}$	–2,1	<0,01

G9W1: Schülerinnen und Schüler aus G9-Jahrgängen der ersten Erhebungswelle; G9W2: Schülerinnen und Schüler aus G9-Jahrgängen der zweiten Erhebungswelle; G8W2: Schülerinnen und Schüler aus G8-Jahrgängen der zweiten Erhebungswelle; G8W3: Schülerinnen und Schüler aus G8-Jahrgängen der dritten Erhebungswelle; G9ges: Schülerinnen und Schüler aller G9-Kohorten; G8ges: Schülerinnen und Schüler aller G8-Kohorten; $G9_{ges}$–$G8_{ges}$: Mittelwertdifferenz aller G9- und G8-Jahrgänge. Die Metrik der latenten Variablen wurde transformiert auf $M = 50$ und $SD = 10$ auf Basis der gepoolten Mittelwerte bzw. Standardabweichungen. Die berichteten p-Werte beziehen sich auf zweiseitige Tests. Subskripte indizieren Unterschiede zwischen den jeweiligen Gruppen. Alle Unterschiede waren signifikant mit $p < 0{,}01$. Es werden nur Unterschiede zwischen den Doppeljahrgängen, den reinen Jahrgängen und Unterschiede innerhalb der G8- und G9-Jahrgänge dargestellt. Die finale Analysestichprobe basiert auf Daten von $N = 4887$ Schülerinnen und Schülern, die am Schülerfragebogen teilgenommen habenn.

achten gilt, dass diese Analysen lediglich die Informationen von Schülerinnen und Schülern berücksichtigen, die Angaben zur durchschnittlichen wöchentlichen Dauer der Aktivitäten in einem Freizeitbereich gemacht haben. Nicht berücksichtigt werden konnte hierbei die relative Betätigungshäufigkeit, da „keine Betätigung" nicht als Antwortoption vorgesehen war und somit nicht von fehlenden Werten („nicht bearbeitet") unterschieden werden konnte. Für den Bereich „Freunde treffen" lag der Unterschied zwischen allen G8- und G9-Schülerinnen und Schülern bei 96 Min. (M_{G9ges}-M_{G8ges}: 95,9, $p < 0{,}01$). Hierbei gaben Schülerinnen und Schüler im G9-Jahrgang durchschnittlich eine längere Beschäftigungsdauer in diesem Freizeitbereich an. Der Unterschied zwischen den beiden reinen G8- und G9-Jahrgängen belief sich hier auf rund 171 Min. (M_{G9W1}-M_{G8W3}: 170,5, $p < 0{,}01$), ebenfalls mit höheren Angaben der G9-Schülerinnen und Schüler. Im Freizeitbereich „Nebenjob" gaben die Schülerinnen und Schüler aus G9-Jahrgängen im Mittel eine höhere zeitliche Investition an als Schülerinnen und Schüler aus G8-Jahrgängen (M_{G9ges}-M_{G8ges}: 75,3, $p < 0{,}01$). Weiterhin fanden sich signifikante Unterschiede für die Bereiche „Sport treiben" und „Fernsehen", die sich auf 18,2 Min. und 22,3 Min. beliefen und bei denen jeweils G9-Schülerinnen und Schüler eine längere Beschäftigungsdauer angaben.

7 Diskussion

Die G8-Reform gilt als *die* zentrale Reform des Gymnasiums des ersten Jahrzehnts im neuen Jahrtausend (Trautwein und Neumann 2008). Mit der vorliegenden Studie konnten nun erstmals – zumindest für ein Bundesland – Befunde vorgestellt werden,

Tab. 5 Bereiche der Freizeitbeschäftigung in Stunden

Bereich	Unadjustierte Ergebnisse					
	$G9W1_a$	$G9W2_b$	$G8W2_c$	$G8W3_d$	$G9_{ges}$	$G8_{ges}$
Freunde treffen	$15,12_{bd}$	$13,27_a$	12,62	$12,09_a$	14,20	12,36**
Nebenjob	$7,92_d$	$7,70_c$	$6,54_b$	$6,29_a$	7,81	6,42**
Sport	$6,68_d$	6,45	6,26	$6,13_a$	6,57	6,20**
Fernsehen	$8,84_d$	8,69	8,37	$8,04_a$	8,77	8,20**
Angebote in Schule	3,16	3,08	3,01	3,05	3,12	3,03
Computer	11,07	10,63	10,80	10,81	10,85	10,0
Lesen	4,64	4,57	4,60	4,37	4,60	4,48
Unternehmungen mit Familie	5,53	5,35	5,59	5,59	5,44	5,59
Orchester	$3,04_b$	$3,49_a$	3,39	3,21	3,27	3,30
Weitere Hobbys	4,46	4,21	$3,97_d$	$4,38_c$	4,34	4,18
Adjustierte Ergebnisse						
Freunde treffen	$14,92_{bd}$	$13,28_a$	$12,93_d$	$12,08_{ac}$	14,10	12,50**
Nebenjob	$7,80_d$	$7,54_c$	$6,68_b$	$6,15_a$	7,67	6,42**
Sport	$6,62_d$	6,43	6,37	$6,08_a$	6,53	6,22*
Fernsehen	$8,83_d$	8,68	8,69	$8,83_a$	8,75	8,38*
Angebote in Schule	3,10	3,09	3,07	3,02	3,10	3,04
Computer	11,01	10,58	10,89	10,82	10,79	10,86
Lesen	4,56	4,47	4,63	4,28	4,52	4,45
Unternehmungen mit Familie	5,51	5,34	5,78	5,57	5,42	5,67
Orchester	$3,05_b$	$3,48_a$	3,29	3,20	3,26	3,25
Weitere Hobbys	4,53	4,19	4,00	4,38	4,36	4,19

G9W1: Schülerinnen und Schüler aus G9-Jahrgängen der ersten Erhebungswelle; G9W2: Schülerinnen und Schüler aus G9-Jahrgängen der zweiten Erhebungswelle; G8W2: Schülerinnen und Schüler aus G8-Jahrgängen der zweiten Erhebungswelle; G8W3: Schülerinnen und Schüler aus G8-Jahrgängen der dritten Erhebungswelle; G9ges: Schülerinnen und Schüler aller G9-Kohorten; G8ges: Schülerinnen und Schüler aller G8-Kohorten. Unadjustierte Modelle wurden hier nicht mit FIML, sondern listwise deletion spezifiziert, da fehlende Werte und „keine Betätigung" im entsprechenden Freizeitbereich nicht getrennt erfasst wurden und somit nicht voneinander trennbar sind. In den adjustierten Modellen wurden lediglich fehlende Werte auf unabhängigen Variablen mittels FIML berücksichtigt. Bezüglich der „Rest"-Kategorie „Sonstiges" ergaben sich für keinen der Gruppenvergleiche statistisch signifikante Unterschiede. Indizes stellen statistisch signifikante Gruppenunterschiede $p < 0,05$ dar. Dargestellt wurden lediglich Unterschiede zwischen den beiden reinen Jahrgängen, den Doppeljahrgängen und Unterschiede innerhalb der G9-, bzw, G8-Jahrgänge. Unterschiede zwischen den gesamten G9- und G8-Jahrgängen wurden mit * gekennzeichnet.
** $p < 0,01$, * $p \leq 0,05$

die auch standardisierte Kompetenzmaße umfassen sowie auf einer repräsentativen Stichprobe beruhen. Im Folgenden werden zunächst die Ergebnisse zusammengefasst und mögliche Erklärungsansätze vorgestellt, bevor auf Implikationen für die Bildungspolitik in Baden-Württemberg sowie dem Bundesgebiet eingegangen wird. Abschließend wird die Rolle der Bildungsforschung bei Bildungsreformen kritisch hinterfragt.

7.1 Zentrale Ergebnisse und Erklärungsansätze

In Bezug auf die Kompetenzen fand sich ein bemerkenswertes Ergebnismuster: Während sich in Mathematik und Physik keinerlei Leistungseinbußen durch G8 fanden, zeigten sich für die Lesekompetenz in Englisch substanzielle sowie für Biologie tendenzielle Unterschiede zugunsten der G9-Absolventen. Eine mögliche Erklärung ist, dass die Umstellung auf G8 in den einzelnen Fächern unterschiedlich gut gelang. Im Fach Englisch kam es in Baden-Württemberg wegen der gleichzeitig zur Umstellung auf G8 erfolgten Einführung des Grundschulenglisch und der parallelen Reduktion des Unterrichtsvolumens in Englisch in der Sekundarstufe I um insgesamt acht Wochenstunden zu einer vorübergehenden Reduktion der Gesamtstundenzahl bis zum Abitur; auch war die Wertigkeit des Faches Englisch wegen des parallelen – inzwischen wieder aufgehobenen – Starts mehrerer Fremdsprachen in Klassenstufe 5 für die Schülerinnen und Schüler ggf. etwas in Frage gestellt. Es könnte auch eine Rolle spielen, dass Englisch nicht nur in der Schule gelernt wird, sondern auch im Freizeitbereich (Fernsehserien, Musik, Reisen, Alltagskultur) eine Rolle spielt und im G9 also auch im nichtschulischen Bereich mehr gelernt werden konnte. Zu beachten ist, dass die Unterschiede im Fach Englisch durchaus substanziell ausfielen; sollten die Absolventen jedoch in nennenswerter Zahl das „gewonnene" Jahr für einen Aufenthalt im englischsprachigen Ausland nutzen, könnte dies den Unterschied rasch ausgleichen.

Prononcierte Unterschiede zugunsten von G9 fanden sich beim schulischen Beanspruchungserleben und den gesundheitlichen Beschwerden. Dies war auch der Fall im Doppeljahrgang. Dieser Befund mag etwas überraschen, da die Schülerinnen und Schüler aus G8 und G9 im Doppeljahrgang gemeinsam die Kurse besuchten und mit exakt denselben schulischen Anforderungen konfrontiert waren. Darüber hinaus zeigten sich Unterschiede zwischen Schülerinnen und Schülern, die ähnlich groß oder leicht größer als die Kohortenunterschiede zwischen G8- und G9-Schülerinnen und Schülern ausfielen. Als Erklärungsansätze kommen deshalb in Frage, dass (1) die Absolventen aus G8 jünger sind und deshalb für dieselben Anforderungen mehr Energie aufwenden müssen, (2) die G8-Absolventen in der Mittelstufe Defizite aufbauten, die in der Oberstufe korrigiert werden, oder (3) die Selbstberichte der G8-Absolventen z. T. auch die öffentliche Diskussion um die erwarteten negativen Folgen der Schulzeitverkürzung widerspiegeln. Leider standen für die Auswertungen keine objektiven Markiervariablen für die Gesundheit zur Verfügung, so dass offen bleiben muss, wie sehr die genannten möglichen Ursachen zu dem Ergebnismuster beigetragen haben. Für die Einordnung der Bedeutsamkeit der Befunde sollte man allerdings darauf hinweisen, dass die Kohortenunterschiede geringer ausfielen als die (in jeder Kohorte auftretenden) Unterschiede zwischen männlichen und weiblichen Abiturienten.

Hinsichtlich der Freizeitaktivitäten bestätigen die vorliegenden Daten die Befürchtungen, wonach es zu einem Einbruch bei „wertvollen" Freizeitaktivitäten bei G8-Absolventen käme, nur sehr bedingt; zum Zeitpunkt des Abiturs fanden sich in der Mehrzahl der berücksichtigten Bereiche keine signifikanten Unterschiede.

Die dokumentierten Befunde entsprechen somit nur teilweise den oftmals vorgebrachten Sorgen in Hinblick auf G8. Die Datenbasis für die hier vorgestellten

Analysen darf hierbei als gut gelten. So wurde im Nationalen Bildungspanel ein Kohorten-Kontroll-Design umgesetzt, bei dem unmittelbar aufeinanderfolgende Kohorten untersucht wurden. Zur Absicherung der Befunde wurde eine Serie von unterschiedlichen Modellen berechnet, die sich in den berücksichtigten Kontrollvariablen unterschieden. Insgesamt zeigten sich hierbei keine oder nur sehr geringe Unterschiede zwischen den adjustierten und den unadjustierten Modellen. Wichtig für eine adäquate Interpretation und Einordnung der Ergebnisse bezüglich der Leistungsunterschiede der Schülerinnen und Schüler ist die Qualität der Messinstrumente. Wie bereits oben angeführt, wurden die Leistungstests latent modelliert, sodass aufgrund der teilweise unbefriedigenden Score-Reliabilität einzelner Instrumente keine Verzerrungen der Effektstärken zu erwarten sind. Darüber hinaus zeigten unsere Analysen durchaus substantielle Zusammenhänge zwischen den Fachnoten am Ende der Sekundarstufe II und den Leistungstests sowie insgesamt geringes DIF (auch in Bezug auf den Kohortenvergleich) und einen moderaten Itemfit. Diese Ergebnisse legen keine aufgabenspezifischen Unterschiede nahe, sondern lassen eher vergleichbare Ergebnisse bei einem größeren Itempool erwarten. Gleichwohl zeigte sich auch, dass das *test targeting* noch nicht vollständig befriedigend war. So ist der Englischtest tendenziell eher leicht für die Schülerinnen und Schüler, während der Physiktest viele schwierige Items enthielt. Diese Tendenz zeigte sich jedoch in gleicher Weise sowohl für Schülerinnen und Schüler aus den G8- als auch den G9-Kohorten. Bezogen auf die Validität der eingesetzten Leistungstests ist zu bemerken, dass diese in unterschiedlichem Ausmaß das Curriculum repräsentieren. Besonders deutlich ist die unvollständige Abdeckung des Curriculums beim Englischtest, der lediglich Lesekompetenz (auf insgesamt eher niedrigem Niveau) erfasst, womit in der vorliegenden Studie beispielsweise der Bereich der produktiven Teilkompetenzen im Englischen nicht berücksichtigt wurde. Wenn man aber davon ausgeht, dass die Leistungstests die kohortenspezifischen Curricula jeweils in vergleichbarer Weise abdecken, dann lassen sich die gefundenen Unterschiede (weitgehend) im Sinne von Effekten der Schulzeitverkürzung auf die jeweils erfasste Kompetenz interpretieren.

7.2 Bildungspolitische Implikationen

Welche Implikationen haben die Ergebnisse in Hinblick auf bildungspolitische Entscheidungen? Sind sie ein Beleg für das Funktionieren von G8 in Baden-Württemberg oder lassen sie sich als Basis für eine Forderung nach Rückkehr zu G9 verwenden? Grundsätzlich ist festzuhalten, dass (1) die Ergebnisse nur einen Teil der Wirkungen von G8 reflektieren und (2) erst durch eine subjektive Gewichtung von Zielen und durch den Vergleich mit Erreichtem mit bildungspolitischen Implikationen angereichert werden (vgl. Bromme et al. 2014). Im vorliegenden Fall dürfte es für eine Abschätzung des „Erfolgs“ der Reform wesentlich darauf ankommen, (1) als wie bedeutsam man die „Kosten“ (also beispielsweise die Kompetenzunterschiede in Englisch und beim Wohlbefinden) von G8 bewertet, (2) ob man annimmt, dass inzwischen vorgenommene Nachregulierungen bei G8 (u. a. Grundschulenglisch sowie Unterstützungsangebote in der Oberstufe) die identifizierten Schwachstellen überwinden und (3) wie positiv man das in G8 „gesparte“ Lebensjahr betrachtet.

Darüber hinaus müssen bei Forderungen nach Wiedereinführungen von G9 nach dem Vorbild von Niedersachsen auch potenzielle ungewollte Nebenwirkungen bedacht werden. So würde eine erneute Reform erstens Ressourcen binden, die – so implizieren es viele empirische Studien – vielleicht effizienter in die Unterrichtsentwicklung investiert werden könnten (z. B. Hattie 2008). Zweitens würde eine Rückkehr zu G9 dafür sorgen, dass es in absehbarer Zeit einen Jahrgang gäbe, bei dem kein Abiturient das allgemeinbildende Gymnasium verlassen würde, was wiederum massive negative Konsequenzen für die Hochschulen des Bundeslandes haben dürfte (ein „Nullerjahrgang" anstatt des „Doppeljahrgangs"). Drittens lässt sich auch spekulieren, ob eine Rückkehr zu G9 angesichts der kürzlich aufgehobenen Verbindlichkeit der Grundschulempfehlungen in Baden-Württemberg eine Veränderung des Schulwahlverhaltens zur Folge haben könnte, was wiederum im Konflikt mit der anvisierten Architektur der Schulformen stehen könnte.

Die Implikationen der vorgelegten Studie beschränken sich nicht auf nur ein Bundesland. Natürlich ist zu berücksichtigen, dass es bundesweit nicht *die* G8-Reform gab – vielmehr kam G8 immer gemeinsam mit bestimmten Veränderungen in der Organisation der Mittelstufe und bestimmten curricularen Veränderungen. In empirischen Studien lassen sich diese zwei Faktoren nur schwer trennen, so dass sich in den Befunden zum Zeitpunkt des Abiturs immer zwei Komponenten, nämlich „G8 plus landesspezifische Regelungen", niederschlagen und die spezifischen Wirkungen nicht generalisierbar sind. Trotzdem hat unsere Studie Implikationen jenseits des lokalen Kontextes eines Bundeslands. So ist festzuhalten, dass es – siehe beispielsweise das Fach Mathematik – sehr wohl möglich ist, auch unter den Bedingungen von G8 das Abitur ohne Qualitätsverlust abzulegen. Zweitens können die identifizierten Unterschiede zwischen den Fächern als (erneuter) Beleg dafür herangezogen werden (vgl. Hattie 2008), dass „äußeren" Faktoren, zu denen auch die Frage von G8 vs. G9 gehört, im Vergleich zur Umsetzung von Qualität im Unterricht eine geringere Rolle spielen.

7.3 Implikationen für Evaluationen bei Reformen

Auf einer abstrakteren Ebene kann die G8-Reform als ein Beleg für die Bedeutungslosigkeit der Erziehungswissenschaft bzw. Bildungsforschung betrachtet werden: Bei der Konzeption der Reform war sie kaum einbezogen und auf begleitende Evaluationsmaßnahmen durch die Wissenschaft, die von Anfang an mit eingeplant hätten werden können, wurde gänzlich verzichtet (vgl. Spiewak 2014). Umgekehrt lässt sich aber auch argumentieren, dass der Verzicht auf die Mitarbeit und Begleitung durch die Erziehungswissenschaft/Bildungsforschung zeigt, wie wichtig diese sein könnte.

So sollten Evaluationen von vornherein mitgeplant werden. Hierbei kann man sowohl an formative (reformbegleitende Erhebungen, die zu unmittelbaren Veränderungen führen können) und summative (die Gesamtwirkung der Reform auf unterschiedliche Kriteriumsmaße prüfende) Elemente denken. Anhand der von uns vorgestellten Studie lässt sich auch aufzeigen, wie das Studiendesign für die summativen Elemente noch aussagekräftiger hätte werden können, wenn die Studie von vornherein als Teil der Reform mitgeplant wird: So wäre es möglich gewesen, Da-

ten auch in der Sekundarstufe I zu sammeln, in der die Beanspruchung durch G8 möglicherweise besonders deutlich ausfällt. Zudem hätten sich in Zusammenarbeit mit den Verantwortlichen im Land zusätzliche Kriteriumsmaße identifizieren und einsetzen lassen, die für (positive und negative) Reformeffekte besonders sensitiv sein könnten.

Natürlich kann und soll eine solche Begleitforschung nicht die politischen Entscheidungen ersetzen oder öffentliche Debatten überflüssig machen. Die Frage beispielsweise, ob der „Gewinn" eines schulfreien Lebensjahres bei G8 es ggf. auch rechtfertigen würde, dass im Abitur gewisse Leistungseinbußen zu verzeichnen sind, und die Frage danach, welcher Zeitaufwand für die Schule gefordert wird und welches Maß an Belastung „akzeptabel" ist, sind normative Entscheidungen, die als Ergebnisse von Aushandlungsprozessen in bildungspolitische Entscheidungen münden. Sie werden nicht von der Bildungsforschung gesteuert – aber diese könnte, wenn man es ihr ermöglicht, entscheidend dazu beitragen, Diskussionsprozesse mithilfe empirischer Befunde zu fundieren (vgl. Bromme et al. 2014).

Literatur

Baumert, J., Bos, W., Klieme, E., Lehmann, R. H., Lehrke, M., Hosenfeld, I., Neubrand, J., & Watermann, R. (Hrsg.). (1999). *Testaufgaben zu TIMSS/III. Mathematisch-naturwissenschaftliche Grundbildung und voruniversitäre Mathematik und Physik der Abschlussklassen der Sekundarstufe II (Population 3)*. Berlin: Max-Planck-Institut für Bildungsforschung.

Bergmüller, S. (2003). Schulische Belastung und gesundheitliche Beschwerden. In G. Haider & C. Reiter (Hrsg.), *PISA 2003. Internationaler Vergleich von Schülerleistungen*. Graz: Leykam.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, *1*(2), 1–11.

Blossfeld, H.-P., Rossbach, H.-G., & Maurice, J. von (Hrsg.) (2011). *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft: Sonderheft 14) Wiesbaden: Springer VS.

Böhm-Kasper, O., & Weishaupt, H. (2002). Belastung und Beanspruchung von Lehrern und Schülern am Gymnasium. *Zeitschrift für Erziehungswissenschaft*, *5*(3), 472–499.

Bromme, R., Prenzel, M., & Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. *Zeitschrift für Erziehungswissenschaft*, *17*(4), 3–54.

Brunner, M., Lang, F. R., & Lüdtke, O. (2014). *Erfassung der fluiden kognitiven Leistungfähigkeit über die Lebensspanne im Rahmen der National Educational Panel Study: Expertise (NEPS Working Paper No. 42)*. Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Büttner, B., & Thomsen, S.L. (2015). Are we spending too many years in school? Causal evidence of the impact of shortening secondary school duration. *German Economic Review, 16*(1), 65–86.

Carroll, J. B. (1963). A model for school learning. *Teachers College Record*, *64*, 723–733.

Carroll, J. B. (1989). The Carroll Model: A 25-Year retrospective and prospective view. *Educational Researcher*, *18*(1), 26–31.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental psychology*, *27*(5), 703–722.

Duchhardt, C. (2015). NEPS Technical Report for Mathematics: Scaling results for the additional study Baden-Wuerttemberg (*NEPS Working Paper No. 59*). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Eberle, F., Gehrer, K., Jaggi, B., Kottonau, J., Oepke, M., & Pflüger, M. (2008). *Evaluation der Maturitätsreform 1995. Schlussbericht zur Phase II*. Bern: Staatssekretariat für Bildung und Forschung SFB.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists (Multivariate applications book series)*. Mahwah, N.J: L. Erlbaum Associates.

Europarat (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen.* Berlin: Langenscheidt.

Fuchs, H.-W. (2004). *Gymnasialbildung im Widerstreit. Die Entwicklung des Gymnasiums seit 1945 und die Rolle der Kultusministerkonferenz*. Frankfurt: Peter Lang.

Ganzeboom, H. B. G., & Treiman, D. J. (2003). Three internationally standardised measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Hrsg.), *Advances in cross-national comparison* (S. 159–193). Boston, MA: Springer US.

Greiner, L., & Himmelrath, A. (2014). Studie zum Turbo-Abi: G8-Stress gibt es gar nicht. http://www.spiegel.de/schulspiegel/abi/studie-zu-turbo-abi-kaum-unterschiede-bei-abiturienten-mit-g8-und-g9-a-986159.html. Zugegriffen: 13. Feb. 2017.

Hattie, J. (2008). *Visible learning: A synthesis of meta-analyses relating to achievement.* London: Routledge.

Heller, K. (2002). *Begabtenförderung im Gymnasium. Ergebnisse einer zehnjährigen Längsschnittstudie*. Opladen: Leske und Budrich.

Herrmann, U. (2002). Achtjähriges Gymnasium? Thesen Pro und Contra. *Die deutsche Schule*, *94*(4), 471–484.

Hübner, N., Rieger, S., & Wagner, W. (2016a). *NEPS technical report for biological competence: Scaling results for the additional study Baden-Württemberg (NEPS survey paper no. 9)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Hübner, N., Rieger, S., & Wagner, W. (2016b). *NEPS technical report for english reading: Scaling results for the additional study Baden-Württemberg (NEPS survey paper no. 10)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Hübner, N., Rieger, S., & Wagner, W. (2016c). *NEPS technical report for physics competence: Scaling results for the additional study Baden-Württemberg (NEPS survey paper no. 11)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Huebener, M., Kuger, S., & Marcus, J. (2016). Increased instruction hours and the widening gap in student performance. *DIW Discussion Paper*, *1561*, 1–42.

IEA DPC (2013). Methodenbericht: NEPS Zusatzstudie zur G8-Reform in Baden-Württemberg. Haupterhebung – Frühjahr 2011 (A72). https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/Methodenbericht_A72.pdf. Zugegriffen: 17. Feb. 2017.

IEA DPC (2014a). Methodenbericht: NEPS Zusatzstudie zur G8-Reform in Baden-Württemberg. Haupterhebung - Frühjahr 2012 (A73). https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/2-0-0/Methodenbericht_A73.pdf. Zugegriffen: 17. Feb. 2017.

IEA DPC (2014b). *Methodenbericht: NEPS Zusatzstudie zur G8-Reform in Baden-Württemberg. Haupterhebung – Frühjahr 2013 (A74)*. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/3-0-0/Methodenbericht_A74.pdf. Zugegriffen: 17. Feb. 2017.

Jacobsen, L., & Buhse, M. (2013). Schulreform: G8 oder G9? Wer will, bleibt länger. http://www.zeit.de/2013/12/G8-Entscheidung-Eltern. Zugegriffen: 13. Feb. 2017.

Kühn, S. M., van Ackeren, I., Bellenberg, G., Reintjes, C., & Brahm, G. im (2013). Wie viele Schuljahre bis zum Abitur? *Zeitschrift für Erziehungswissenschaft*, *16*(1), 115–136.

Kultusministerium Baden-Württemberg (2004a). Bildungsplan 2004. Allgemein bildendes Gymnasium. http://www.bildung-staerkt-menschen.de/service/downloads/Bildungsplaene/Gymnasium/Gymnasium_Bildungsplan_Gesamt.pdf. Zugegriffen: 13. Feb. 2017.

Kultusministerium Baden-Württemberg (2004b). Gymnasium 2004. Das pädagogische Konzept. http://www.kultusportal-bw.de/site/pbs-bw/get/documents/KULTUS.Dachmandant/KULTUS/import/pb5start/pdf/Gymnasium%202004%20Das%20pdagogische%20Konzept%20G8_Sept2004_klein.pdf. Zugegriffen: 13. Feb. 2017.

Kultusministerium Niedersachsen (2014). Fragen und Antworten zum modernen Abitur nach 13 Jahren. Kultusministerium Niedersachsen. www.mk.niedersachsen.de/download/85662/Fragen_und_Antworten_zum_modernen_Abitur_nach_13_Jahren_hier_herunterladen.pdf. Zugegriffen: 13. Feb. 2017.

Kultusministerkonferenz (2014). Die gymnasiale Oberstufe. http://www.kmk.org/bildung-schule/allgemeine-bildung/sekundarstufe-ii-gymnasiale-oberstufe.html. Zugegriffen: 30. Sep. 2014.

Laging, R., Böcker, P., & Dirks, F. (2014). Zum Einfluss der Schulzeitverkürzung (G8) auf Bewegungs- und Sportaktivitäten von Jugendlichen. *Sportunterricht*, *63*(3), 66–72.
Landesinstitut für Schulentwicklung. (1999). Stundentafel für die Klassen 5 bis 11 des allgemein bildenden Gymnasiums: Sprachliches Profil und naturwissenschaftliches Profil. http://www.ls-bw.de/,Lde/Startseite/Bildungsplaene/sprachliches+u_+naturwissenschaftliches+profil. Zugegriffen: 17. Feb. 2017.
Milde-Busch, A., Blaschek, A., Borggräfe, I., von Kries, R., Straube, A., & Heinen, F. (2010). Besteht ein Zusammenhang zwischen der verkürzten Gymnasialzeit und Kopfschmerzen und gesundheitlichen Belastungen bei Schülern im Jugendalter? *Klinische Pädiatrie*, *222*(4), 255–260.
Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2013). Ministerin Löhrmann: NRW hat Abitur mit Doppeljahrgang erfolgreich bewältigt: Ergebnisse des Zentralabiturs 2013. https://www.schulministerium.nrw.de/docs/bp/Ministerium/Presse/Pressekonferenzen/2013/130822Zentralabitur/Zentralabitur_Sprechzettel-1.pdf. Zugegriffen: 17. Feb. 2017.
Morgan, S.L., & Winship, C. (2007). *Counterfactuals and causal inference. Methods and principles for social research (Analytical methods for social research)*. New York: Cambridge University Press.
Murnane, R.J., & Willett, J.B. (2011). *Methods matter. Improving causal inference in educational and social science research*. Oxford: Oxford University Press.
Muthén, B., & Muthén, L.K. (2012). *Mplus User's Guide* (7. Aufl.). Los Angeles: Muthén & Muthén.
NEPS (2011). G8-Reform in Baden-Württemberg: Haupterhebung 2010/11 (A72) Schüler/innen, Klasse 12/13 Informationen zum Kompetenztest. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/2-0-0/C_A72_de.pdf. Zugegriffen: 13. Feb. 2017.
Patall, E.A., Cooper, H., & Allen, A.B. (2010). Extending the school day or school year: A systematic review of research (1985–2009). *Review of Educational Research*, *80*(3), 401–436.
Reckase, M.D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
Quis, J. S. (2015). *Does higher learning intensity affect student well-being? Evidence from the National Educational Panel Study.* BERG Working Paper Series, 94.
Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.
Rupp, A.A., & Zumbo, B.D. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement*, *66*(1), 63–84.
Rupp, A.A., Vock, M., Harsch, C., & Köller, O. (2008). *Developing standards-based assessment tasks for english as a first language. Context, processes and outcomes in Germany. Bd. 1: Standards-Based Assessment Tasks for English as a First Language.* Münster: Waxmann.
Samejima, F. (1997). Graded response model. In W.J. van der Linden & R.K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 85–100). New York, NY: Springer.
Scheerens, J. (Hrsg.) (2014). *Effectiveness of time investments in education*. Cham: Springer.
Schönberger, B., & Aßmann, C. (2014). Weighting the Additional Study in Baden-Wuerttemberg of the National Educational Panel Study, National Educational Panel Study. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/2-0-0/BW_2-0-0_Weighting.pdf. Zugegriffen: 17. Feb. 2017.
Schul-Volksbegehren in Niedersachsen (2011). Turbo-Abitur wird Wahlkampfthema. http://www.taz.de/!5121786/. Zugegriffen: 13. Feb. 2017.
Schwarz-Jung, S. (2008). Allgemeinbildende Gymnasien in Baden-Württemberg flächendeckend fünf Jahrgänge im „G8". *Statistisches Monatsheft Baden-Württemberg*, *10*, 3–10.
Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
Spiewak, M. (2014). Glaubenskrieg um vier Stunden. Das Hickhack um G8 oder G9 ist eine Armutserklärung für die Bildungspolitik. Nun rufen Schulforscher Halt. http://www.zeit.de/2014/25/g8-debatte-bildungspolitik. Zugegriffen: 13. Feb. 2017.
Statistische Ämter des Bundes und der Länder (2015). D13.1 und D13.2: Anzahl und Anteile der Klassenwiederholungen. http://www.statistikportal.de/Statistik-Portal/. Zugegriffen: 13. Feb. 2017.
Statistisches Landesamt Baden-Württemberg (2014a). Sommer 2013: Rund 15500 Schülerinnen und Schüler an Werkreal-/Hauptschulen, Realschulen und Gymnasien verfehlen das Klassenziel. http://www.statistik-bw.de/Pressemitt/2014257.asp. Zugegriffen: 13. Feb. 2017.
Statistisches Landesamt Baden-Württemberg (2014b). Übergänge aus Klassenstufe 4 an Grundschulen auf weiterführende Schulen (öffentliche und private Schulen). http://www.statistik-portal.de/BildungKultur/Landesdaten/uebergaenger_0000.asp?y=2003. Zugegriffen: 13. Feb. 2017.

Trautwein, U., & Neumann, M. (2008). Das Gymnasium. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer, & L. Trommer (Hrsg.), *Das Bildungswesen in der Bundesrepublik Deutschland* Rororo Sachbuch, (Bd. 62339, S. 467–501). Reinbek bei Hamburg: Rowohlt.

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*(4), 788–806.

Trautwein, U., Lüdtke, O., Becker, M., Neumann, M., & Nagy, G. (2008). Die Sekundarstufe I im Spiegel der empirischen Bildungsforschung: Schulleistungsentwicklung, Kompetenzniveaus und die Aussagekraft von Schulnoten. In E. Schlemmer & H. Gerstberger (Hrsg.), *Ausbildungsfähigkeit im Spannungsfeld zwischen Wissenschaft, Politik und Praxis* (S. 91–107). Wiesbaden: Springer VS.

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O., & Maaz, K. (Hrsg.) (2010). *Schulleistungen von Abiturienten. Die neugeordnete gymnasiale Oberstufe auf dem Prüfstand*. Wiesbaden: Springer VS.

Tulodetzki, P., & Gohr, L. (2012). Turbo-Abi. „Die Schule wird uns stressiggeredet". http://www.focus.de/familie/schule/turbo-abi/die-schule-wird-uns-stressiggeredet-turbo-abi_id_2454177.html. Zugegriffen: 13. Feb. 2017.

Vieth-Entus, S. (2014). G8, G9 und Turboabitur. Das größte Problem sind die Kultusminister. http://www.tagesspiegel.de/meinung/g8-g9-und-turboabitur-das-groesste-problem-sind-die-kultusminister/9840140.html. Zugegriffen: 13. Feb. 2017.

Weiler, H. N. (2003). Bildungsforschung und Bildungsreform – Von den Defiziten der deutschen Erziehungswissenschaft. In I. Gogolin & R. Tippelt (Hrsg.), *Innovation durch Bildung. Beiträge zum 18. Kongress der Deutschen Gesellschaft für Erziehungswissenschaft* (S. 181–203). Opladen: Leske & Budrich.

Zumbo, B. (1999). *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.